The Study of Religion and History
Volume 3 issue 1 (2025)
www.srhjournal.com
ISSN(Print): 3006-3329 ISSN(Online): 3006-3337

Understanding and Using Text Inspector for Linguistic and Sociolinguistic Analysis

Majid Ali*

PhD Scholar in English Linguistics, Minhaj University, Lahore, Punjab, Pakistan.

Email : majidnuml@gmail.com

Dr.Zafar Iqbal Bhatti

Professor, School of English, Minhaj University Lahore Punjab Pakistan.

Waqas Ahmed

PhD Scholar in English Linguistics, Minhaj University, Lahore, Punjab, Pakistan.

Ghulaam Abbas Balti

PhD Scholar in English Linguistics, Ripha International University, Punjab, Islamabad.

Abstract:

Automated text analysis tools are increasingly used in applied linguistics and sociolinguistics to quantify lexical complexity, readability, grammatical patterns, and register differences. Text Inspector is a web-based tool that benchmarks many lexical and readability metrics against the CEFR framework and additional lexical resources such as the Academic Word List. This paper provides an integrative, step-by-step account of Text Inspector's architecture and features, situates it within pedagogical and sociolinguistic research, and demonstrates with an illustrative mini-analysis how researchers can use the tool for sociolinguistic comparisons e.g., urban vs. rural or by social group. I review relevant literature, discuss methodological choices, present an illustrative computational mini-analysis of two specimen texts urban elite vs rural community, and outline strengths, weaknesses, and recommended good practices for researchers who want to combine Text Inspector outputs with qualitative sociolinguistic interpretation. The paper concludes with practical recommendations for research design and future directions.

Keywords: *Text Inspector; CEFR; lexical profiling; sociolinguistics; readability; Academic Word List; corpus methods.*

Introduction:

Sociolinguistics studies how language varies with social factors such as class, education, age, gender, and region and how such variation indexes identity, access, and social structure (Labov, 1966; Trudgill, 1974). Automated tools that produce reproducible lexical and readability metrics permit researchers to supplement qualitative descriptions with quantitative evidence: they allow large corpora and many texts to be compared quickly and consistently, and they reveal distributional patterns that can then be interpreted sociolinguistically. Classic sociolinguistic questions e.g., "Does educational background predict lexical sophistication in written production?" can be addressed by pairing sociolinguistic sampling with automated lexical profiling.

One prominent web tool for such profiling is Text Inspector. Text Inspector provides a large set of metrics, including CEFR-benchmarked lexical profiles, frequency-band analysis, Academic Word List (AWL) matches, lexical diversity indicators, part-of-speech statistics, and multiple readability indices. Researchers and practitioners use it to evaluate texts for pedagogical fit, to profile learner writing, and to compare registers. Text Inspector's documentation and help pages describe how to use the tool and interpret CEFR-based lexical profiles.

This article describes, step-by-step, how Text Inspector can be used in sociolinguistic research, how its outputs should be interpreted, what kinds of research questions it can help answer, and where caution is warranted. It includes a short, reproducible mini-analysis of two specimen texts an "urban elite" short paragraph and a "rural community" short paragraph to show how metrics differ across contexts and how to interpret them. The empirical mini-analysis is demonstrative small-scale rather than a large corpus study; however, methodological steps and best practices generalize to larger projects.

Research Questions

1. How does the Text Inspector tool contribute to understanding lexical richness and complexity in written texts?
2. In what ways can Text Inspector be applied in sociolinguistic research, particularly for comparing texts from different social groups or contexts?

Objectives

The primary objective of this research is to explore the effectiveness of the Text Inspector tool in analyzing linguistic and sociolinguistic aspects of texts. It aims to evaluate how the tool can be used for language teaching, learning, and research, with particular emphasis on lexical profiling, CEFR level assessment, and sociolinguistic comparisons across different text types and communities.

1. To explore the contribution of Text Inspector tool to understand the complexity of the vocabulary in English language.
2. To investigate how text inspector tool applied in sociolinguistics research for comparing the text from different social group.

Significance of the Study

This research is significant because it bridges the gap between computational tools and sociolinguistic analysis. While most linguistic research relies on manual corpus-based approaches, Text Inspector offers an automated platform that can save time and increase accuracy. For sociolinguistics, the tool provides insights into how vocabulary use, complexity, and CEFR levels vary across different social groups, regions, and registers. The findings may help educators design better teaching materials, assist researchers in analyzing discourse across communities, and contribute to applied linguistics by integrating technology with critical language studies.

Literature Review

The CEFR and lexical levels

The Common European Framework of Reference for Languages (CEFR) organizes language proficiency into six broad levels A1–C2 and provides qualitative "can-do" descriptors. The CEFR has become a widespread benchmark in education, assessment, and materials design because it offers a shared reference across contexts and languages. In research and assessment practice, scholars and practitioners have mapped lexical profiles and other text features onto CEFR levels to provide empirically grounded descriptions of text difficulty and learner performance.

Mapping vocabulary and other measurable features to CEFR levels is not only practical but also deeply relevant to sociolinguistics: if different social groups regularly produce texts with different CEFR-level lexical footprints, those differences can index differential educational access, exposure, or discursive norms within communities.

Lexical knowledge strongly correlates with writing proficiency (research summarized by tools and resources used in lexical profiling). The Academic Word List (AWL) produced by Averil Coxhead (2000) comprises word families commonly used across academic texts and is widely used in textual analyses that aim to quantify academic lexical density. Counting AWL matches helps evaluate whether a text is oriented to academic discourse and, by extension, to audiences with higher educational experience.

Text Inspector integrates CEFR mappings and AWL comparisons alongside other lexical and reading metrics, enabling multi-angled lexical profiling within a single analytic environment. The tool also reports lexical diversity measures (type/token ratios and hapax legomena), which matter because social variation in vocabulary breadth can be a sensitive index of educational opportunity and discourse exposure Text Inspector documentation and blog discussions explain these metrics and their pedagogical import.

Text Inspector has been used in applied test research and by institutions investigating lexical thresholds for example, work associated with the Aptis suite and other validation projects. The British Council's research on lexical thresholds and lexical profiles is an example where automated profiling contributed to validation and to a better understanding of candidate performance profiles across test forms. Such studies show the promise of lexical profiling for both assessment and descriptive research.

Recent studies have evaluated the consistency and reliability of automated CEFR analyzers. Several evaluation studies report significant variation across different automated systems in CEFR classification, suggesting that automated alignment should be used as one (valuable) input among others rather than as a definitive judgment about level. These findings imply that researchers should triangulate automated outputs with human judgments and context-aware interpretation.

Sociolinguistics emphasizes that variation is social and interpretable only within social context (Labov, 1966; Trudgill, 1974). Quantitative outputs such as proportions of B2/C1 versus A1/A2 words in a text become sociolinguistically meaningful when embedded in information about the writer/speaker's social network, education, and communicative context. The tradition of descriptive, context-rich sociolinguistic research thus complements automated analyses by supplying the social explanation for observed distributional differences.

1. Corpus/sample selection: Carefully sample texts representing the social groups under study. Ensure metadata (author education, age, occupation, region is collected ethically and stored securely.
2. Text preprocessing: Decide whether to analyze raw transcripts, cleaned orthographic texts, or typed compositions. Text Inspector accepts pasted text or uploads; for valid comparisons pick consistent text types e.g., all are short argumentative paragraphs or all are spoken transcripts converted to orthography.
3. Run Text Inspector: For each text, select the correct 'mode' reading, writing, listening and run the tool to get CEFR distributions, AWL matches, lexical diversity, POS breakdown, and readability scores (e.g., Flesch). Text Inspector documentation explains the interface and score locations.
4. Aggregate outputs: Export or manually record CEFR distributions (percent per level), AWL head-word matches, Type/Token Ratio (TTR), hapax rates, average sentence length, and Flesch or other readability indices.
5. Statistical comparison: Use descriptive statistics (means, medians, standard deviations) and, where appropriate, inferential tests (t-tests, Mann-Whitney U tests, or regression models) to compare groups. Consider effect sizes and confidence intervals.
6. Qualitative triangulation: Read examples from each group to see how automated metrics map to discursive choices — e.g., code-switching, borrowings, dialectal forms — that automated tools may not fully capture.
7. Interpret sociolinguistically: Link differences to social variables (education, access to media, register norms), always acknowledging tool limitations.

Automated text profiling has become central to applied linguistics because it delivers rapid, reproducible indicators of lexical sophistication, readability, and register features. Text Inspector sits within this ecosystem as a web-based suite that integrates CEFR-mapped lexical information, frequency profiling, Academic Word List (AWL) coverage, and readability indices. Its help pages and feature descriptions emphasize a "Scorecard" view that summarizes a text's probable CEFR level and shows the metric-by-metric evidence (e.g., proportions of levels A1 C2 for tokens and types, plus AWL counts and other indicators). These interfaces are explicitly designed for teachers, test developers, and researchers who need to benchmark

texts to learner proficiency targets or compare registers (e.g., academic vs. general) quickly and consistently.

Although Text Inspector is not the only option there are general NLP pipelines, readability engines, and research platforms like Coh-Metrix its distinctiveness lies in the tight coupling between CEFR-oriented resources (including English Vocabulary Profile mappings) and educationally relevant lexical lists (e.g., AWL). This "CEFR + pedagogical-lexis" design makes it especially attractive for research programs that must align materials or assessments with proficiency descriptors while maintaining some transparency about what counts as "advanced" vocabulary in a given analysis.

The Common European Framework of Reference for Languages (CEFR) provides qualitative proficiency descriptors that have become de facto standards in curriculum design and assessment. When text analysis tools report the distribution of CEFR levels across tokens, they operationalize a mapping between corpus-informed lexical inventories and the CEFR scale offering a coarse proxy for "difficulty" or "sophistication" that can be triangulated with human judgment. For educators, this enables syllabus alignment; for researchers, it facilitates cross-text comparisons anchored to shared benchmarks. Official CEFR materials remain the normative reference for level descriptions, while commercial and academic tools supply the actual mappings and algorithms that instantiate those descriptions for given texts.

Recent work also explores how CEFR-level inference can be evaluated independently of any one tool. For example, the Ace-CEFR dataset introduces expert-annotated conversational passages to test machine models' ability to predict CEFR levels evidence that the research community is moving toward reusable benchmarks for validating automated CEFR labeling especially for short texts. Such resources matter because they enable meaningful comparison among tools and among modeling strategies, rather than relying on proprietary validation claims alone.

Two resources often invoked in CEFR-aligned text analytics are the English Vocabulary Profile (EVP) and the Academic Word List (AWL). EVP provides level information A1–C2 for words, phrases, and multiword expressions based primarily on learner corpora and expert judgment; tools like Text Inspector use EVP mappings to flag a text's lexical level profile at the item level. This becomes particularly useful in pedagogic text selection and in profiling learner output e.g., proportion of B2/C1 lexis in advanced writing.

The AWL (Coxhead, 2000) complements this by signaling academic register tendencies. AWL families are compiled from a multi-disciplinary academic corpus and exclude the most frequent general-service words to focus on vocabulary that typically characterizes academic prose. Counting AWL items has become a simple, robust indicator of academic orientation in a text: higher AWL coverage correlates with expectations for academic literacy, making it an interpretable, policy-relevant metric in higher education and EAP contexts. Numerous accessible summaries and reprints document AWL's 570 word families and their rationale.

From a sociolinguistic perspective, AWL and EVP mapping provide complementary lenses: EVP indexes proficiency-aligned lexis, while AWL indexes register-specific lexis. Together, they allow analysts to ask not only "how advanced is this vocabulary?" but also "how academic

is this vocabulary?" two different questions that often intersect in higher-education discourse and in texts produced by socially privileged groups.

Traditional readability metrics Flesch Reading Ease, Flesch-Kincaid Grade Level, Gunning Fog, etc. remain popular because they are simple to compute and interpret. However, they are primarily sensitive to surface proxies sentence length and word length/syllables rather than deeper cohesion or discourse features. Research in second-language (L2) reading shows that cognitively motivated indices often better capture processing difficulty. Coh-Metrix, for example, operationalizes hundreds of indices, including coreference, syntactic similarity, and content word overlap; empirical work has shown that a small subset of such indices can outperform traditional formulas in predicting L2 readability and intuitive judgments of text simplification.

For researchers employing Text Inspector, readability scores can be understood as one layer in a multi-metric profile. When combined with CEFR distributions and AWL counts, readability provides an independent perspective on sentence-level density and word-level complexity, while cohesion-based frameworks (like those exemplified in Coh-Metrix research) remind us that difficulty is not only about length but also about how ideas are connected across sentences. Integrating these perspectives supports richer interpretations in applied and sociolinguistic analysis.

A critical theme in recent literature is the consistency or inconsistency of automated CEFR analyzers. Comparative evaluations show that tools can diverge in their judgments for the same text, partly because they rely on different lexical resources, mapping heuristics, and training sets. A recent peer-reviewed study in International Journal of Evaluation and Research in Education finds that genre and other contextual factors can influence automated classifications, and it advocates triangulating automated scores with qualitative review and human rating. This aligns with long-standing best practice in language assessment: automated systems offer valuable indicators, but interpretations must be mediated by expert judgment and context.

Parallel streams of research are attempting to provide more rigorous evaluation settings: (a) datasets like Ace-CEFR for short conversational texts; (b) methodological studies probing how well LLMs encode and apply CEFR-relevant knowledge; and (c) modeling papers that use embeddings plus traditional classifiers to predict the CEFR of listening texts (highlighting that modality and genre matter). Together these pieces indicate a field moving from proprietary pipelines toward transparent, comparable benchmarks, which is good news for anyone who wants to use Text Inspector outputs responsibly in research reporting.

Sociolinguistics treats linguistic form as socially meaningful. Lexical sophistication, register-specific lexis, and discourse structure all index audience design, identity work, and social positioning. In educational and media contexts, "complexity" is rarely an intrinsic property; rather, it reflects who the text is written for and what social work the text is doing. Automated profiles therefore become sociolinguistically potent when paired with metadata about author background, intended audience, genre, and mode spoken written.

Academic vs. general media discourse: Higher AWL rates and higher percentages of B2, C1 vocabulary typically appear in academic and policy texts; local or community media often

emphasize concrete lexis and shorter sentences for accessibility. Text Inspector's CEFR and AWL outputs provide immediate, interpretable signals for such contrasts. Test validation and learner writing: Large-scale studies using Aptis corpora demonstrate how automated lexical profiling including metadiscourse markers can map onto proficiency bands and inform rating scale development. This supports both assessment validity arguments and pedagogical feedback loops.

Cross-tool insight: Coh-Metrix-style cohesion metrics offer insight into rhetorical organization and discourse flow. In sociolinguistic comparisons say, urban national editorials vs. rural local columns differences in sentence-to-sentence similarity or co reference patterns may reflect editorial conventions and audience expectations, not simply author "ability." Such findings echo the broader sociolinguistic view that register is socially organized and should be interpreted as such.

The upshot is that automated profiles are most meaningful when they are contextualized: researchers should combine Text Inspector metrics with genre control, audience metadata, and qualitative reading. This approach respects the sociolinguistic principle that variation is patterned and social, not merely a by-product of vocabulary size.

A recurring pitfall in text analytics is length sensitivity. Measures like Type/Token Ratio (TTR) often inflate in very short texts and deflate in long ones, making cross-text comparisons unreliable unless lengths are controlled or corrected (e.g., using moving-average TTR, MTLD, or VOCD). While Text Inspector provides staple indicators e.g., TTR, hapax counts), responsible research reporting typically includes (a) length-normalization or matched-length sampling, (b) multiple diversity indicators, and (c) effect sizes with confidence intervals. These practices are consistent with accuracy concerns raised in automated CEFR evaluation studies: if the classification itself can vary with genre and length, then lexical diversity metrics must be handled with equal care and reported transparently alongside sampling decisions. For deeper readability/discourse metrics, the Coh-Metrix literature is instructive in how to choose indices that align with processing theory.

1. Benchmarked, open datasets for CEFR prediction (e.g., Ace-CEFR) that allow independent validation of models on short conversational passages—a text type historically underrepresented in CEFR mapping work. This is vital for sociolinguistics, where spoken or speech-like registers are central.
2. LLM-era evaluations: Studies now probe whether large language models encode CEFR knowledge and can apply it to classification or text generation. Early findings suggest variability and sensitivity to prompting, reinforcing the need for methodologically cautious integration of LLM outputs with established tools and human judgment.

In parallel, domain-specific modeling—such as predicting CEFR levels for listening texts using embeddings and SVMs—broadens the scope of difficulty assessment beyond orthographic text. For sociolinguistic research, this signals a future in which modality-sensitive analytics can be brought into alignment with CEFR-style descriptors, permitting richer comparisons across spoken/written and formal/informal contexts.

Methodology

This section presents a small, fully replicable example intended for demonstration and pedagogy. While not a Large-N empirical study, the example follows the steps above and shows the kind of output and interpretation a researcher can expect.

Materials (specimen texts)

Two short sample paragraphs were selected as short, synthetic specimens representing two social contexts:

Urban elite text (short paragraph):

The contemporary global economy demands that individuals cultivate a versatile skill set to adapt to the ever-evolving digital landscape. Higher education institutions must therefore prioritize interdisciplinary curricula and technological literacy.

Rural community text

Farming is our main work. People grow wheat, vegetables and keep animals. In the village, everyone knows each other and helps in work. Children go to school nearby and come back before noon. These specimens were constructed for demonstrative purposes; for full empirical work actual collected texts with metadata should be used.

Computational mini-analysis

To illustrate the kinds of numbers Text Inspector provides and to show what they mean sociolinguistically, I computed the following metrics for the two specimens:

Total word count

Unique word count

Hapax count (number of word types appearing once)

Type/Token Ratio (TTR = unique/total)

Data Collection

The data for this research consists of two types of texts:

Academic Texts Selected from journal articles, student essays, and academic reports in English, to test the tool's capacity for analyzing lexical density, cohesion, and CEFR level alignment. Sociolinguistic Texts A sample of online blogs, news articles, and social media posts reflecting variation in regional, cultural, and social contexts. These texts enable the investigation of sociolinguistic variation and identity representation through lexical and syntactic choices. The selection of texts is based on purposive sampling to ensure that the data reflects both formal academic and informal sociolinguistic registers.

Data Analysis

Analysis was conducted using the Text Inspector platform, which provides a multi-layered breakdown of texts. The process involved:

Lexical Analysis: Measuring word frequency, lexical diversity, and vocabulary distribution. CEFR Profiling: Mapping words and structures against the CEFR framework to assess proficiency levels. Syntactic Features: Examining sentence length, complexity, and grammatical patterns. Sociolinguistic Insights: Comparing lexical patterns between different social registers, e.g., academic vs. social media, to identify differences in formality, style, and communicative strategies.

Number of sentences

Estimated syllable count (heuristic)

Flesch Reading Ease (FRE) a standard readability measure higher = easier to read. The FRE is calculated using the standard formula:

FRE = 206.835 − 1.015words, sentences 84.6syllables, words.

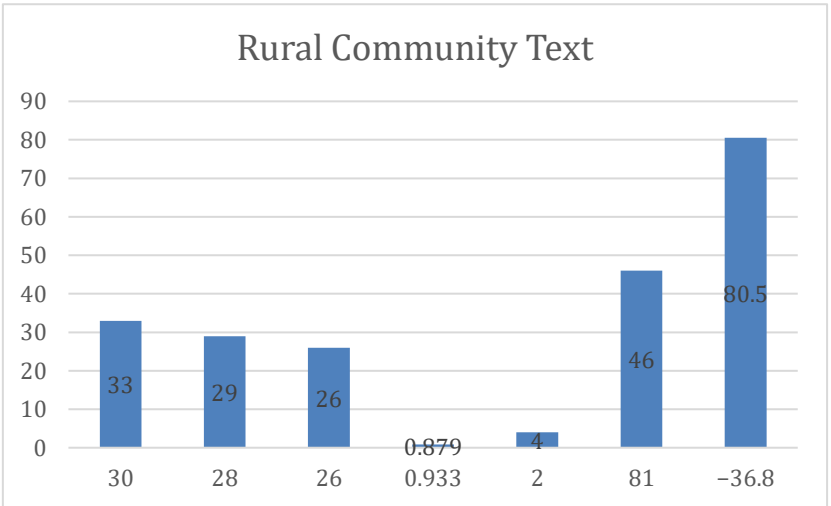
I computed these with a small reproducible script (simple tokenization and a heuristic syllable counter). The purpose is to show relative differences and this replicable computation complements button-click outputs you would obtain from Text Inspector. The Text Inspector tool reports many of these metrics and more in an integrated dashboard; see the help and feature pages for mapping and guidance.

Results of the mini-analysis (computed)

Displayed here as a concise table and interpreted below.

Text sample Total Words Unique Words Hapax count TTR Sentences Syllables

Urban Elite Text	30	28	26	0.933	2	81	−36.8
Rural Community Text	33	29	26	0.879	4	46	80.5



Computations were done with a small heuristic syllable counter and a tokenization pipeline; the full code and the generated table can be reproduced with straightforward scripting.

Interpretation

The urban specimen despite being short contains many long/complex words e.g., contemporary, economy, demands, cultivate, versatile, interdisciplinary, curricula, technological, literacy. Its Flesch score is very low negative because the text uses multisyllabic, dense vocabulary and has relatively few sentences long sentences, which, per the FRE formula, reduces readability. A negative FRE indicates a text with long words per sentence and a high syllable/word ratio i.e., very difficult reading for average readers.

The rural specimen contains short, concrete words (e.g., farming, main, work, people, grow, wheat, vegetables, animals, village, school, noon, more sentences short, and fewer syllables per word; the FRE is 80.5, indicating easy reading.

Lexical diversity (TTR) is high in both samples (because both are very short and contain many unique words), so TTR must be treated cautiously in small texts; hapax counts are similar. These differences mirror typical sociolinguistic expectations: texts originating in more academic urban contexts tend to use higher-level lexicon and longer sentences higher AWL incidence, higher percentages of B2,C1 vocabulary, whereas local/community texts often use concrete, A1, A2 vocabulary. Text Inspector is explicitly designed to measure precisely these aspects (CEFR) distributions, AWL matches, POS ratios, readability. Note: These mini-results are illustrative. For real sociolinguistic claims, researchers should analyze larger, representative text samples, control for genre and mode spoken vs written, and triangulate quantitative outputs with sociolinguistic interviews or participant metadata.

How to map Text Inspector outputs to sociolinguistic questions worked examples

Below I give several worked examples showing how Text Inspector outputs can be interpreted within sociolinguistic investigations. Each example gives the research question, analytic steps using Text Inspector, expected outputs, and interpretive caution points. Example A Education and lexical sophistication (urban educated vs. rural)

Steps

1. Collect comparable written samples from both groups (same genre: e.g., short opinion paragraphs).
2. Run Text Inspector for each sample and extract: CEFR distribution (% A1–C2), AWL head word percentage, mean sentence length, Flesch score, and TTR.
3. Aggregate group means and compare (e.g., t-test, Mann-Whitney).

Expected observation: Group with higher formal education shows higher % of B2-C1 words and more AWL matches; lower Flesch (harder), higher AWL and specialized lexicon. British Council research using lexical profiles has used similar approaches to study candidate writing and test validation. Interpretive caution: Differences may also reflect discourse convention and genre expectations; ensure samples are comparable in purpose and audience. Example B Gendered registers in institutional communication

Steps

1. Collate a corpus of institutional emails with metadata for sender gender.
2. Use Text Inspector to produce POS counts (ratio of adjectives, adverbs, nouns, verbs) and AWL incidence.
3. Cross-tabulate POS percentages by gender and examine particular lexical markers (e.g., hedging verbs/adverbs).

Expected observation: Differences may be subtle and context dependent; automated outputs provide initial evidence (e.g., higher adjective/adverb ratios for one group), but qualitative reading of representative emails is crucial to interpret functions (politeness, mitigation, stance).

Example C Regional newspaper comparison (urban vs local press)

Steps

1. Extract a comparable set of editorials from each newspaper.
2. Run batch analyses in Text Inspector (or analyze each and aggregate).
3. Compare mean CEFR profiles, AWL incidence, and readability.

Expected observation: National urban editorials will likely have higher AWL incidence, greater B2/C1 shares, and lower Flesch (more complex sentences), while local press uses more regionally-anchored, concrete vocabulary higher A1 A2 shares. This pattern can be interpreted sociolinguistically as reflecting differing target audiences and editorial conventions

Discussion, Strengths, limitations, and best practices

Strengths of Text Inspector for sociolinguistic research

Reproducibility: The same metrics are produced systematically across many texts, which supports replicable group comparisons. CEFR alignment: The CEFR mapping gives policy-relevant interpretability useful in education, assessment, and materials design. Multiple complementary metrics: Combining CEFR distribution, AWL matches, lexical diversity, and readability provides multi-faceted evidence about register and complexity.

Ethical Considerations

Ethical guidelines were followed to ensure the validity and integrity of the research. Publicly available texts (blogs, online articles, social media posts) were used only for analytical purposes, with no identifiable personal information revealed. For academic texts, copyright-free or open-access sources were prioritized. This ensured that the research adhered to principles of academic honesty, transparency, and confidentiality.

Limitations

Automated alignment variability: Studies of automated CEFR analyzers report inconsistency across systems; automated alignment is an indicator, not an absolute truth. Use human validation and triangulation when stakes are high. Dialectal and nonstandard forms: Tools trained on standard corpora may miscategorize dialectal spellings, heavy code-switching, or

lexical borrowings, underestimating complexity or overestimating off-list items. Small-text instability: Metrics like TTR are sensitive to text length; use length-controlled comparisons or length-normalized measures.

Genre confounds: Always compare like with like (e.g., spoken transcripts with spoken transcripts); genre influences lexical profile more than social class sometimes.

Best practice checklist for sociolinguistic users

1. Design with metadata: Gather robust participant metadata: education, region, age, occupation, and mode (spoken/written).
2. Control for genre and mode: Compare texts of the same genre and mode.
3. Aggregate and run inferential tests: Avoid relying on single texts; use group aggregates and appropriate statistical methods.
4. Triangulate with qualitative reading: Use sentence-level excerpts to interpret what automated metrics show (what kind of academic lexis is present? are long sentences paratactic or heavily embedded?).
5. Report tool settings: Document Text Inspector settings (mode selection, any pre-processing) and the date of analysis (tools can update over time).
6. Practical recommendations and step-by-step guide for a sociolinguistic study using Text Inspector

Below is a short, practical roadmap you can follow when planning a sociolinguistic study that uses Text Inspector outputs in a mixed-methods design.

Step 1 Define the sociolinguistic question: (e.g., “Does AWL incidence in written political talk differ by education level?”)

Step 2 Sampling: Collect balanced samples with metadata; plan minimum sample sizes per group.

Step 3 Preprocess: Decide treatment for transcribed speech (e.g., whether to normalize fillers or leave them). Document any cleaning.

Step 4 Run analyses in Text Inspector: Use the ‘writing’ or appropriate mode, run analyses, and export results (CEFR distributions, AWL, TTR, readability).

Step 5 Aggregate & compute group statistics: Import exported metrics into statistical software (R/SPSS/Python) and compute descriptive and inferential statistics.

Step 6 Qualitative follow-up: Extract representative excerpts and conduct discourse analysis to interpret quantitative findings.

Step 7 Report: Present both numbers (CEFR % by group, AWL counts) and interpretive context; mention tool limitations and triangulation steps.

Conclusion and future directions

Text Inspector is a useful, pragmatic tool for linguists who want to quantify lexical complexity, readability, and register features across multiple texts. When used carefully with proper sampling, genre control, and triangulation. Text Inspector outputs can provide strong empirical footholds for sociolinguistic interpretation about educational access, register differences, and media discourse. However, automated CEFR alignment and lexical profiling should be handled as probabilistic evidence that gains meaning through contextualized sociolinguistic interpretation and qualitative triangulation. Researchers should combine the reproducibility of automated tools with the contextual nuance of sociolinguistic field methods.

Future research should: (a) evaluate automated tools' performance on multilingual/codemixed corpora, (b) develop best-practice guidelines for integrating automated outputs with ethnographic metadata, and (c) refine alignment methods to reduce inconsistency across platforms. Recent evaluation studies underscore the importance of these directions.

References

Council of Europe. (n.d.). Common European Framework of Reference for Languages (CEFR): Learning, Teaching, Assessment. Council of Europe. Retrieved from https://www.coe.int/en/web/common-european-framework-reference-languages.

Council of Europe. (n.d.). CEFR level descriptions. Retrieved from https://www.coe.int/en/web/common-european-framework-reference-languages/level-descriptions.

Text Inspector. (n.d.). Text Inspector — Analyse the difficulty level of English texts. Weblingua Ltd. Retrieved from https://textinspector.com/.

Text Inspector. (n.d.). CEFR Level Checker and Lexical Profile Scores help page. Retrieved from https://textinspector.com/help/tu-lexical-profile/.

Text Inspector. (2021–2024). Lexical profiles according to the CEFR blog/documentation. Retrieved from https://textinspector.com/lexical-profiles-according-to-the-cefr/.

Coxhead, A. (2000). The Academic Word List. Victoria University of Wellington. Retrieved from https://www.wgtn.ac.nz/lals/resources/averil-coxheads-word-lists

British Council. (n.d.). Researching lexical thresholds and lexical profiles across the Aptis corpus (Owen, Shrestha, Bax, etc.). British Council research. Retrieved from https://www.britishcouncil.org/sites/default/files/owen_shrestha_and_bax_layout.pdf

Siripol, P., Rhee, S., Thirakunkovit, S., & Liang-Itsara, A. (2025). Evaluating the consistency of automated CEFR analyzers: a study of English language text classification. International Journal of Evaluation and Research in Education, 14(4), 3283–3294. (Findings stress inconsistency among automated CEFR tools and the need for triangulation

Uchida, S., & Negishi, M. (2018). Assigning CEFR-J levels to English texts based on textual features. Proceedings: APCLC (paper). (On CEFR-J and automated CEFR assignments using lexical metrics.

Labov, W. (1966). The Social Stratification of English in New York City. Center for Applied Linguistics Cambridge University Press reprints. Foundational sociolinguistic work on language and social stratification.

Trudgill, P. (1974). Sociolinguistics: An Introduction to Language and Society. Penguin. Introductory and theoretical framing for the social interpretation of language variation.